

# APPROXIMATION ALGORITHMS FOR THE NORMALIZING CONSTANT OF GIBBS DISTRIBUTIONS

BY MARK HUBER

Consider a family of distributions  $\{\pi_\beta\}$  where  $X \sim \pi_\beta$  means that  $\mathbb{P}(X = x) = \exp(-\beta H(x))/Z(\beta)$ . Here  $Z(\beta)$  is the proper normalizing constant, equal to  $\sum_x \exp(-\beta H(x))$ . Then  $\{\pi_\beta\}$  is known as a Gibbs distribution, and  $Z(\beta)$  is the partition function. This work presents a new method for approximating the partition function to a specified level of relative accuracy using only a number of samples that is  $O(\ln(Z(\beta)) \ln(\ln(Z(\beta))))$  when  $Z(0) \geq 1$ . This is a sharp improvement over previous similar approaches, which used a much more complicated algorithm requiring  $O(\ln(Z(\beta)) \ln(\ln(Z(\beta))))^5$  samples.

**1. Introduction.** The central idea of Monte Carlo methods is that the ability to sample from certain distributions gives a means for estimating the value of an integral or sum. This paper presents a new method for using samples to approximate a broad class of sums coming from Gibbs distributions that is faster than previously known methods.

**DEFINITION 1.1.**  $\{\pi_\beta\}_{\beta \in \mathbb{R}}$  is a *Gibbs distribution with parameter  $\beta$*  over finite state space  $\Omega$  if there exists a *Hamiltonian function*  $H(x) : \Omega \rightarrow \mathbb{R}$  such that for  $X \sim \pi_\beta$ ,

$$\mathbb{P}(X = x) = \exp(-\beta H(x))/Z(\beta),$$

where  $Z(\beta) = \sum_{x \in \Omega} \exp(-\beta H(x))$  is called the *partition function* of the distribution.

The partition function can be difficult to compute, even when dealing with simple problems.

**EXAMPLE 1.1** (The Ising model). Given a graph  $G = (V, E)$ , let  $\Omega = \{-1, 1\}^V$ , and  $H(x) = -\sum_{\{i,j\} \in E} \mathbf{1}(x(i) = x(j))$ , where  $\mathbf{1}(\cdot)$  is the indicator function that is 1 if the argument is true and 0 if it is false. Then the Gibbs distribution with this Hamiltonian is called the *Ising model*. Finding  $Z(\beta)$  for arbitrary graphs is a #P-complete problem [8].

---

*AMS 2000 subject classifications:* Primary 68Q87, 65C60; secondary 65C05

*Keywords and phrases:* integration, Monte Carlo methods, cooling schedule, self-reducible

A vast literature has arisen devoted to finding ways to generate random variables from Gibbs distributions (see for instance [9, 12] or [2] for an overview.) For the Ising model, Jerrum and Sinclair [8] gave an algorithm for approximately sampling from  $\pi_\beta$  in polynomial time for any value of  $\beta$ . Propp and Wilson [10] gave an algorithm for the Ising model that seems to run efficiently when  $\beta$  is at or below a cutoff known as the critical value.

Once an effective method for obtaining approximate or perfect samples from the target Gibbs distribution exists, the question becomes, what is the best way of using those samples to approximate  $Z(\beta)$ ?

DEFINITION 1.2. Say that  $\mathcal{A}$  is an  $(\epsilon, 3/4)$ -*randomized approximation algorithm* for  $Z(\beta)$  if it outputs value  $\hat{Z}(\beta)$  such that

$$\mathbb{P}\left(\frac{1}{1+\epsilon} \leq \frac{\hat{Z}(\beta)}{Z(\beta)} \leq 1+\epsilon\right) \geq 3/4.$$

Here  $\epsilon \geq 0$  controls the relative error between the approximation and the true answer. The  $3/4$  on the right hand side can be made arbitrarily close to 1 by repeating the algorithm and taking the median of the resulting output.

1.1. *Previous work.* The first step in building such an approximation algorithm is importance sampling. For most Gibbs distributions calculating  $Z(0)$  is straightforward, and it is easy to generate samples from  $\pi_0$ . For the Ising model,  $\pi_0$  is just the uniform distribution over  $\{-1, 1\}^V$ , and  $Z(0) = 2^{\#V}$ . With a draw  $X \sim \pi_0$  in hand, let

$$(1.1) \quad W = \exp(-\beta H(X)).$$

Then

$$\mathbb{E}[W] = \frac{\sum_{x \in \Omega} \exp(-\beta H(x)) \exp(0)}{Z(0)} = \frac{Z(\beta)}{Z(0)},$$

making  $W \cdot Z(0)$  an unbiased estimator of  $Z(\beta)$ .

The relative performance of this Monte Carlo estimate is controlled by the relative variance, the square of the coefficient of variation. For a random variable  $X$  with finite second moment,  $\mathbb{V}_{\text{rel}}(X) = [\mathbb{E}(X^2)/\mathbb{E}(X)^2] - 1$ . Hence for the random variable  $W$  as in (1.1):

$$(1.2) \quad \mathbb{V}_{\text{rel}}(W) = -1 + \frac{\sum_{x \in \Omega} \exp(-\beta H(x))^2}{Z(0)} \cdot \frac{Z(0)^2}{Z(\beta)^2} = -1 + \frac{Z(2\beta)Z(0)}{Z(\beta)^2}.$$

There are two main issues with this relative variance:

1. For problems like the Ising model, this last ratio can be exponentially large in the input, making the method untenable.
2. The relative variance involves the value of  $Z(2\beta)$ , outside the interval of interest  $[0, \beta]$ . Typically, larger values of  $\beta$  make sampling from  $\pi_\beta$  more difficult. This presents a serious impediment to the method.

The first problem can be dealt with by using the *multistage sampling* method of Valleeau and Card [14]. In this approach, a sequence of  $\beta$  values  $0 = \beta_0 < \beta_1 < \beta_2 < \dots < \beta_\ell = \beta$  are introduced, called a *cooling schedule*. Then

$$\frac{Z(\beta)}{Z(0)} = \frac{Z(\beta_1)}{Z(\beta_0)} \cdot \frac{Z(\beta_2)}{Z(\beta_1)} \cdots \frac{Z(\beta_\ell)}{Z(\beta_{\ell-1})}.$$

Each of the individual factors in the product on the right can then be estimated separately and then multiplied to give a final estimate. Fishman called an estimate of this form a *product estimator* [5, p. 437].

It is simple to calculate the mean and relative variance of a product estimator in terms of the mean and relative variance of the individual factors.

**THEOREM 1.1.** *For  $P = \prod P_i$  where the  $P_i$  are independent,*

$$\mathbb{E}[P] = \prod \mathbb{E}[P_i], \quad \mathbb{V}_{rel}(P) = -1 + \prod (1 + \mathbb{V}_{rel}(P_i)).$$

Let  $q = \ln(Z(\beta)/Z(0))$ , and suppose  $H(x) \in \{0, \dots, n\}$ . Then Bezáková et al. [1] introduced a fixed cooling schedule with two pieces, the first where the parameter value grows linearly, and the second where it grows exponentially:

$$0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{k}{n}, \frac{k\gamma}{n}, \frac{k\gamma^2}{n}, \dots, \frac{k\gamma^t}{n}$$

where  $k = \lceil q \rceil$  and  $\gamma = 1 + 1/q$ . With this fixed cooling schedule, they give an  $(\epsilon, 3/4)$ -approximation algorithm that uses  $O(q^2(\ln n)^2)$  samples in the worse case.

By using an adaptive cooling schedule, it is possible to do better. In [7], the author and Schott introduced a general technique for finding normalizing constants of sums and integrals called TPA. When applied to the specific problems of Gibbs distributions, the running time for an  $(\epsilon, 3/4)$ -approximation algorithm becomes  $O(q^2)$ .

To break the  $q^2$  barrier for Gibbs distributions, Štefankovič, Vempala and Vigoda [13] created a multistage algorithm that adaptively created the cooling schedule. The resulting algorithm was highly complex, and they were interested primarily in the theoretical properties rather than a practical implementation. Their  $(\epsilon, 3/4)$ -approximation algorithm used at most

$$(1.3) \quad 10^8 q(\ln(n) + \ln(q))^5 \epsilon^{-2}$$

samples on average from the target distribution.

**1.2. Main result.** The multistage idea solves the issue of  $Z(2\beta)Z(0)/Z(\beta)^2$  being too large, but fails to solve the issue of the variance depending on  $Z(2\beta)$ . Dealing with this leads to several of the  $\ln$  factors in [13]. In this work a new method is introduced, the *paired product estimator* which has a variance only involving quantities within  $[0, \beta]$ . The result is an algorithm where the overall variance can be analyzed precisely. This allows for the construction of an approximation algorithm much simpler than that found in [13], and which requires far fewer samples.

**THEOREM 1.2.** *Suppose  $n \geq 4$  and  $\epsilon \leq 1/10$ . When  $H(x) \in \{0, 1, \dots, n\}$  or  $\{0, -1, -2, \dots, -n\}$ , the new method is an  $(\epsilon, 3/4)$ -approximation algorithm that uses only*

$$(1.4) \quad (q+1)[5 + (2 + \ln(2n))(14.9 \ln(100(2 + \ln(2n))(q+1)) + 48.2\epsilon^{-2})].$$

*draws from the Gibbs distribution on average. In the more general case that  $|H(X)| \leq n$ , the new method is an  $(\epsilon, 3/4)$ -approximation algorithm that uses at most*

$$(q + 2n\beta + 1)[5 + 10.7 \ln(69.4(q + 2n\beta + 1)) + 16.7\epsilon^{-2}].$$

*draws from the Gibbs distribution on average.*

It is possible to derive an upper bound on the number of samples used when  $n < 4$  or  $\epsilon > 1/10$ , these assumptions just make the presentation cleaner. The first statement of the theorem is proved in section 3, and the second statement is proved in section 4.2.

Section 2 describes the overall structure of the algorithm and shows how to obtain a good cooling schedule. Section 3 then analyzes the relative variance of the pieces of the algorithm. The next section then describes how to build a randomized approximation algorithm that uses the number of samples given in (1.4). Section 4 then considers some variations and extensions of the algorithm and how the number of samples needed changes.

**2. The Algorithm.** Let  $q = \ln(Z(0)/Z(\beta))$ . Then to obtain an approximation within a factor of  $1 + \epsilon$  of  $Z(0)/Z(\beta)$ , it is necessary to obtain an approximation of  $q$  within an additive factor of  $\ln(1 + \epsilon)$ . The main algorithm consists of the following pieces.

1. Obtain an initial estimate of  $q$ .
2. Obtain a well balanced cooling schedule.

3. Use the well balanced schedule with the paired product estimator.

The first two pieces will be accomplished using TPA, introduced in [7]. To use TPA for Gibbs distributions on parameter values  $[0, \beta]$ , it is necessary that  $H(x)$  be either always nonnegative or always positive. In section 4.2 it is shown how to deal with the situation where the sign of  $H(x)$  changes over the state space, but for now assume that the sign is unchanging.

In the Ising model example shown earlier,  $H(x) \leq 0$ , and so  $Z(\beta)$  is an increasing function of  $\beta$ . In this case, TPA is an algorithm that generates a random set of parameter values in the interval from 0 to  $\beta$  by taking samples from  $\pi_b$  for various values of  $b \in [0, \beta]$ . The output of TPA is a Poisson point process (PPP) of rate 1 in  $[z(0), z(\beta)]$ . (See section 2 of [7].)

**ALGORITHM 2.1.** TPA for Gibbs distributions with  $H(x) \leq 0$  takes as input a value  $\beta > 0$  together with an oracle for generating random samples from  $\pi_b$  for  $b \in [0, \beta]$ , and returns a set of values  $0 < b_1 < b_2 < \dots < b_\ell < \beta$  such that  $\{z(b_1), \dots, z(b_\ell)\}$  forms a Poisson point process of rate 1 on the interval  $[z(0), z(\beta)]$ . It operates as follows.

1. Start with  $b$  equal to  $\beta$  and  $B$  equal to the empty set.
2. Draw a random sample  $X$  from  $\pi_b$ , and draw  $U$  uniformly from  $[0, 1]$ .
3. Let  $b = b - \ln(U)/H(X)$ , unless  $H(X) = 0$ , in which case set  $b = -\infty$ .
4. If  $b > 0$ , then add  $b$  to the set  $B$ , and go back to step 2.

The number of samples drawn by TPA will equal 1 plus a Poisson random variable with mean  $q$  [7, pp. 3–4]. The output of Algorithm 2.1 can be used in several different ways. When TPA is run  $k$  times and the output sets combined, the result is a Poisson point process on  $[z(0), z(\beta)]$  of rate  $k$ .

It is even possible to obtain rates that are fractional. To obtain rate  $k$  where  $k$  is not an integer, first run TPA  $\lceil k \rceil$  times. Then for each point of the process, keep it independently with probability  $k/\lceil k \rceil$ . Otherwise discard it entirely. This procedure, known as *thinning*, enables creation of a PPP of any positive rate, which will simplify the analysis later. (See [11, p. 320] for more on thinning.)

After a PPP of rate  $k$  has been generated, the number of points in the process has a Poisson distribution with mean  $k(z(\beta) - z(0))$ . This gives a way of initially getting an estimate of  $z(\beta) - z(0)$  that (by choosing  $k$  high enough) has a 99% chance of being within a factor of 2 of the correct value.

Once that is accomplished, TPA is run, this time with an even larger value of  $k$  based on the estimate from the first step. Because the  $z(b)$  values form a Poisson point process, the difference between successive  $z(b)$  values will be an exponential random variable, so if  $b'$  is the  $d$ th point following  $b$ , then

$z(b') - z(b)$  will have a gamma (Erlang) distribution with shape parameter  $d$  and rate parameter  $k$ . By making  $k$  and  $d$  large enough, this will be tightly concentrated around its mean value of  $d/k$  for all such differences.

The result is a set of parameter values  $\{\beta_i\}$  that are *well balanced* in the sense that for these  $z(\beta_{i+1}) - z(\beta_i)$  are all close to the same small value.

Call  $[\beta_i, \beta_{i+1}]$  interval  $i$ . Now each  $z(\beta_{i+1}) - z(\beta_i)$  will be estimated independently using the paired product estimator. This works as follows. For each interval  $i$ , let  $m_i = (\beta_i + \beta_{i+1})/2$  be the midpoint of the interval, and  $\delta_i = m_i - \beta_i = \beta_{i+1} - m_i$  be the half length of an interval. Draw  $X \sim \pi_{\beta_i}$ , and  $Y \sim \pi_{\beta_{i+1}}$ . Then set

$$W_i = \exp(-\delta_i H(X)), \quad V_i = \exp(\delta_i H(Y))$$

Then

$$\mathbb{E}[W_i] = \frac{\sum \exp(-\beta_i H(x)) \exp(-\delta_i H(x))}{Z(\beta_i)} = \frac{\sum \exp(-m_i H(x))}{Z(\beta_i)} = \frac{Z(m_i)}{Z(\beta_i)}.$$

Similarly,  $\mathbb{E}[V_i] = Z(m_i)/Z(\beta_{i+1})$ . Therefore,  $W_i$  can be used to estimate the drop  $z(m_i) - z(\beta_i)$ , and  $V_i$  can estimate the drop  $z(\beta_{i+1}) - z(m_i)$ .

Now for the relative variance calculation.

$$\begin{aligned} \mathbb{V}_{\text{rel}}(W_i) &= \frac{\mathbb{E}[W_i^2]}{\mathbb{E}[W_i]^2} - 1 = -1 + \frac{\sum \exp(-\beta_i H(x)) \exp(-\delta_i H(x))^2}{Z(\beta_i)} \cdot \frac{Z(\beta_i)^2}{Z(m_i)^2} \\ &= -1 + \frac{Z(\beta_{i+1})Z(\beta_i)}{Z(m_i)^2} \text{ since } \beta_i + 2\delta_i = \beta_{i+1}. \end{aligned}$$

A similar calculation shows that  $\mathbb{V}_{\text{rel}}(V_i) = \mathbb{V}_{\text{rel}}(W_i)$ , and now the variance of our estimators for interval  $i$  only involves  $Z(b)$  values for  $b$  that fall in interval  $i$ .

Let  $W$  be the product of the  $W_i$  over all intervals  $i$ , and  $V$  be the product of the  $V_i$ . Then the final estimate of  $Z(\beta)/Z(0)$  is  $W/V$ . This is not quite an unbiased estimator, but it is true that  $\mathbb{E}[W]/\mathbb{E}[V] = Z(\beta)/Z(0)$ . If both  $W$  and  $V$  are tightly concentrated around their means, then  $W/V$  will be close to  $Z(\beta)/Z(0)$ . To get that tight concentration, in the next section it is shown that the relative variance of  $W$  (and  $V$ ) is small as long as the  $\beta$  values form a well balanced schedule.

With that small relative variance it is possible to repeatedly draw independent, identical copies of  $W$  to get a sample average  $\bar{W}$  which is tightly concentrated about its mean. (The same is true for  $V$  as well.) The following algorithm incorporates these ideas.

**ALGORITHM 2.2** (Paired product approximation algorithm). The input is a value  $\beta > 0$  together with an oracle for generating samples from  $\pi_b$  for  $b \in [0, \beta]$ . The output is an approximation for  $Z(\beta)/Z(0)$ .

1. Run TPA 5 times to get an estimate of  $q = \ln(Z(\beta)/Z(0))$  that is at least  $q/2$  with probability 99%.
2. Run TPA  $k$  times to obtain a set of parameter values. Sort these values and then keep every  $d$  successive value. Add parameter values 0 and  $\beta$  and label the result  $0 = \beta_0 < \beta_1 < \dots < \beta_\ell = \beta$ .
3. Repeat the following  $\lceil 2e\sqrt{10}((1+\epsilon)^{1/2} - 1)^{-2} \rceil$  times: for each  $i$ , draw  $X_i \sim \pi_{\beta_i}$ , let  $W_i = \exp(-\delta_i H(X_i))$  and  $V_i = \exp(\delta_i H(X_{i+1}))$ ,  $W = \prod W_i$  and  $V = \prod V_i$ . Take the sample average of the  $W$  values to get  $\bar{W}$ , and the sample average of the  $V$  values to get  $\bar{V}$ .
4. The estimate of  $Z(\beta)/Z(0)$  is  $\bar{W}/\bar{V}$ .

Note that  $((1+\epsilon)^{1/2} - 1)^{-2} \approx 4\epsilon^{-2}$ . It is necessary to use this more complex expression because the final estimator is the ratio of  $W$  and  $V$  (see the proof of Theorem 3.2.) Algorithm 2.2 can be run for any values of  $d$  and  $k$ , the next section shows how to choose them properly to make Algorithm 2.2 an  $(\epsilon, 3/4)$ -approximation algorithm.

**3. Analysis.** In this section the following theorem is shown.

**THEOREM 3.1.** *In Algorithm 2.2, let  $\hat{q}_1$  be the size of the Poisson point process created with 5 runs of TPA in step 1. Let*

$$d = \lceil 22 \ln(100(2 + \ln(2n))(\hat{q}_1 + 1/2)) \rceil, \text{ and } k = (2/3)d[2 + \ln(2n)].$$

*Then the algorithm output is within  $1 + \epsilon$  of  $Z(\beta)/Z(0)$  with probability at least  $3/4$ .*

Let  $q = \ln(Z(\beta)/Z(0))$ . The proof breaks into three parts. The first shows that by running TPA 5 times, the probability that  $\hat{q}_1 + 1/2 < (1/2)q$  is at most 1%. The second part shows that with the choice of  $k$ , the probability that the schedule is not well balanced is at most 4%. Finally, the third part shows that the third step of the algorithm produces  $\bar{W}$  and  $\bar{V}$  that are both within  $1 + \tilde{\epsilon}/2$  of their respective means with probability at most 20%. The union bound on the probability of failure is then  $1\% + 4\% + 20\% = 25\%$ , as desired.

**3.1. The initial estimate  $\hat{q}_1$ .** Recall that Algorithm 2.1 has output that is a Poisson point process with rate 1. Let  $k_1$  denote the number of times that

TPA is run and the output combined. Then the new PPP has a rate of  $k_1$ . Therefore the number of points in the PPP is Poisson distributed with mean  $k_1(z(\beta) - z(0))$ . The following lemma concerning Poisson random variables then shows that  $\hat{q}_1 + 1/2$  is at least  $1/2$  of its mean with probability at least 99%.

LEMMA 3.1. *Let  $X$  have Poisson distribution with mean  $\mu$ . Then  $\mathbb{P}(X < \mu/2) \leq 2(\pi\mu)^{-1/2}(2/e)^{\mu/2}$ .*

PROOF. Suppose  $\mu/2 = \lceil \mu/2 \rceil$ . Then

$$\mathbb{P}(X < \mu/2) = \exp(-\mu) \sum_{i \leq \mu/2} \frac{\mu^i}{i!} \leq \exp(-\mu) 2 \frac{\mu^{\mu/2}}{(\mu/2)!}.$$

The last inequality comes from the fact that each term in the sum is at least twice the previous term. The Sterling bound  $i! > \sqrt{2\pi i}(i/e)^i$  gives  $\mathbb{P}(X \leq \mu/2) \leq 2(\pi\mu)^{-1/2}(2/e)^{\mu/2}$ . Now suppose  $\mu/2 \neq \lceil \mu/2 \rceil$ . Let  $\mu' = 2\lceil \mu/2 \rceil$ .

$$\mathbb{P}(X < \mu/2) \leq \mathbb{P}(X \leq \mu'/2) \leq 2(\pi\mu')^{-1/2}(2/e)^{\mu'/2} \leq 2(\pi\mu)^{-1/2}(2/e)^{\mu}.$$

□

Suppose step 1 runs  $k_1$  repetitions of TPA. Then  $\hat{q}_1$  has a Poisson distribution with mean  $k_1 q$ . If  $q \leq 1$  then it is always true that  $\hat{q}_1 + 1/2 \geq (1/2)q$ . If  $q > 1$  then setting  $k_1 = 5$  and using Lemma 3.1 makes the probability of failure below 1%.

3.2. *The well balanced schedule.* Now consider the second step in Algorithm 2.2. First, run TPA  $k$  times to get a set  $B$  that is a PPP of rate  $k$  on the interval  $[z(0), z(\beta)]$ . Since  $B$  is a PPP of rate  $k$ , if  $b < b'$  are values in  $B$  such that there are exactly  $d - 1$  values in  $(b, b')$ , then  $z(b') - z(b)$  has a gamma distribution with parameters  $d$  and  $k$ . That is equivalent to saying  $z(b') - z(b)$  has the distribution of the sum of  $d$  independent exponential random variables each with rate  $k$ . Hence the moment generating function of  $z(b') - z(b)$  is  $[k/(k - t)]^d$ . Let  $t$  and  $\eta$  be nonnegative real numbers, then

$$\begin{aligned} \mathbb{P}(z(b') - z(b) \geq \eta) &= \mathbb{P}(\exp(t(z(b') - z(b))) \geq \exp(\eta t)) \\ &= [k/(k - t)]^d \exp(-\eta t) \text{ by Markov's inequality} \\ &= (\eta k/d)^d \exp(-\eta k + d) \text{ by setting } t = k - d/\eta. \end{aligned}$$



On the other hand, for  $t > 0$ , multiplying by  $-t$  and exponentiating gives

$$\begin{aligned}\mathbb{P}(z(b') - z(b) \leq \eta/2) &= \mathbb{P}(\exp(-t(z(b') - z(b))) \geq \exp(-\eta t/2)) \\ &= [k/(k+t)]^d \exp(\eta t/2) \text{ by Markov's inequality} \\ &= (\eta k/(2d))^d \exp(-\eta k/2 + d) \text{ by setting } t = 2d/\eta - k.\end{aligned}$$

So if  $d = (3/4)\eta k$ , then from the union bound

$$\mathbb{P}(\eta/2 \leq z(b') - z(b) \leq \eta) \geq 1 - [\exp(-1/3) \cdot 4/3]^d - [\exp(1/3) \cdot 2/3]^d.$$

For the PPP, the chance that  $z(b) - z(b') \in [\eta/2, \eta]$  for the first  $2\eta^{-1}(z(\beta) - z(0))$  intervals to the left of  $\beta$  is (again by the union bound) at least  $1 - 2\eta^{-1}(z(\beta) - z(0))2[\exp(-1/3) \cdot 4/3]^d$ . Making

$$d \geq \frac{\ln(0.04(4\eta^{-1}(z(\beta) - z(0)))^{-1})}{-(1/3) + \ln(4/3)} = \frac{\ln(100\eta^{-1}(z(\beta) - z(0)))}{1/3 - \ln(4/3)}$$

would make this probability at least 96%. However,  $q = z(\beta) - z(0)$  is unknown. What is known (from step 1 of Algorithm 2.2) is  $2(\hat{q}_1 + 1/2)$  has a 96% chance of being at least  $q$ . Since  $(1/3 - \ln(4/3))^{-1} = 21.905\dots$ , setting

$$d = \lceil 22 \ln(200\eta^{-1}(\hat{q} + 1/2)) \rceil$$

and  $k = (4/3)d/\eta$  makes the chance that step 2 fails to find a schedule where  $z(b) - z(b') > 1$  for any interval at most 4%.

**3.3. Choosing  $\eta$ .** The next question to consider is the size of  $\eta$ . The value of  $\eta$  will be used to control the overall relative variance of the product estimators  $W$  and  $V$ . For the  $i$ th interval  $[\beta_i, \beta_{i+1}]$ , let  $m_i \stackrel{\text{def}}{=} (\beta_i + \beta_{i+1})/2$  be the midpoint of the interval. Let  $\delta_i$  be the difference between the  $y$ -coordinate of the midpoint of the interval secant line and the function value at the midpoint of the interval. That is,

$$\delta_i \stackrel{\text{def}}{=} \frac{z(\beta_{i+1}) + z(\beta_i)}{2} - z(m_i).$$

From (1.2),  $\mathbb{V}_{\text{rel}}(W_i) = \exp(2\delta_i) - 1$ . Since the relative variance is always nonnegative, this implies that  $\delta_i \geq 0$  and so the function  $z$  is convex.

From Theorem 1.1,

$$(3.1) \quad \mathbb{V}_{\text{rel}}(W) = -1 + \prod (1 + \exp(2\delta_i) - 1) = -1 + \exp\left(\sum 2\delta_i\right).$$

So controlling the overall relative variance is a matter of bounding  $\delta_i$  for each interval  $i$ . The key idea in the bound comes from [13], although they use it in a very different fashion. The idea is that when  $\delta_i$  is large, the derivative of  $z$  sharply increases.

LEMMA 3.2. *For the  $i$ th interval  $[\beta_i, \beta_{i+1}]$  with  $z(\beta_{i+1}) - z(\beta_i) = \eta_i$ ,*

$$\frac{z'(\beta_{i+1})}{z'(\beta_i)} \geq \exp(4\delta_i/\eta_i).$$

PROOF. Let  $m_i = (\beta_i + \beta_{i+1})/2$  be the midpoint of interval  $i$ , and  $\eta_i = z(\beta_{i+1}) - z(\beta_i)$  be the change in the  $z$  function over the interval. Since  $z$  is convex, the slope at  $\beta_i$  is at most  $[z(m_i) - z(\beta_i)]/[m_i - \beta_i]$ . On the other hand, the slope at  $\beta_{i+1}$  is at least  $[z(\beta_{i+1}) - z(m_i)]/[\beta_{i+1} - m_i]$ . Since  $m_i$  is the midpoint of the interval,  $m_i - \beta_i = \beta_{i+1} - m_i$ , and

$$\frac{z'(\beta_{i+1})}{z'(\beta_i)} \geq \frac{z(\beta_{i+1}) - z(m_i)}{z(m_i) - z(\beta_i)} = \frac{\eta_i/2 + \delta_i}{\eta_i/2 - \delta_i} = \frac{1 + 2\delta_i/\eta_i}{1 - 2\delta_i/\eta_i} \geq \exp(4\delta_i/\eta_i).$$

□

LEMMA 3.3. *For a cooling schedule over  $[0, \beta]$  with  $z(\beta_{i+1}) - z(\beta_i) \leq \eta$  for all  $i$ ,*

$$\mathbb{V}_{\text{rel}}(W) = \mathbb{V}_{\text{rel}}(V) \leq 2e^\eta [2z'(\beta)]^{\eta/2}.$$

*In particular, for  $n \geq 4$  and  $\eta = 2/[2 + \ln(2n)]$ ,*

$$\mathbb{V}_{\text{rel}}(W) = \mathbb{V}_{\text{rel}}(V) \leq 2e.$$

PROOF. Recall that  $\mathbb{V}_{\text{rel}}(W) \leq \exp(2 \sum_i \delta_i)$  so the goal is to bound  $\sum_i \delta_i$ .

Consider a cooling schedule  $0 = \beta_0 < \beta_1 < \dots < \beta_\ell = \beta$ . It is well known that  $z'(\beta)$  is just  $\mathbb{E}[-H(X)]$  where  $X \sim \pi_\beta$ :

$$z'(\beta) = \frac{d}{d\beta} \ln(Z(\beta)) = \frac{Z'(\beta)}{Z(\beta)} = \frac{\sum_x -H(x) \exp(-\beta H(x))}{Z(\beta)} = \mathbb{E}[-H(X)].$$

Case I:  $z'(\beta) < 1/2$ . Then  $H(x) \leq -1 \Rightarrow -H(x) \geq 1$  so

$$\begin{aligned} \frac{\sum_{x: H(x) \leq -1} -H(x) \exp(-\beta H(x))}{Z(\beta)} &\leq \frac{1}{2} \Rightarrow \frac{\sum_{x: H(x) \leq -1} \exp(-\beta H(x))}{Z(\beta)} \leq \frac{1}{2} \\ &\Rightarrow \frac{\sum_{x: H(x)=0} \exp(-\beta H(x))}{Z(\beta)} \geq \frac{1}{2} \\ &\Rightarrow \frac{Z(0)}{Z(\beta)} \geq \frac{1}{2}. \end{aligned}$$

Hence  $z(\beta) - z(0) \leq \ln(2)$  which means  $\sum_i 2\delta_i \leq \ln(2)$  and  $\exp(\sum_i 2\delta_i) \leq 2$ .

Case II:  $z'(0) \geq 1/2$ . Then  $2z'(\beta) \geq z'(\beta)/z'(0)$ , and from the last lemma

$$\frac{z'(\beta)}{z'(0)} = \frac{z'(\beta_1)}{z'(\beta_0)} \dots \frac{z'(\beta_\ell)}{z'(\beta_{\ell-1})} \geq \prod_i \exp(4\delta_i/\eta_i).$$

Raising to the  $\eta/2$  power then finishes this case.

Case III:  $z'(0) < 1/2 \leq z'(\beta)$ . Since  $z'$  is continuous, let  $a \in [0, \beta]$  be the parameter value where  $\mathbb{E}[-H(X)] = 1/2$  for  $X \sim \pi_a$ , and suppose  $a$  is in the  $j$ th interval  $[\beta_j, \beta_{j+1}]$ . As in Case I,  $Z(\beta_j)/Z(\beta_0) \leq 2$ . As in Case II,  $\prod_{i>j} \exp(4\delta_i) \leq [2z'(\beta)]^{\eta/2}$ . Since  $2\delta_j \leq \eta$ , that means the combined relative variance is at most  $2e^\eta [2z'(\beta)]^{\eta/2}$ .

Since  $z'(\beta) = \mathbb{E}[-H(X)]$  for  $X \sim \pi_\beta$ , and  $X \leq n$ ,  $z'(\beta) \leq n$ . Hence if  $\eta/2 \leq 1/[2 + \ln(2n)]$ , then  $e^\eta [2z'(\beta)]^{\eta/2} \leq e$ .  $\square$

PROOF OF THEOREM 3.1. Using the value of  $d$  from section 3.2 and Lemma 3.3 gives that the relative variance for an instance of  $W$  (or  $V$ ) is at most  $2e$ . All that remains is to analyze the third step of Algorithm 2.2. It is easy to verify that if  $\bar{W}$  is the sample average of  $r$  independent, identically distributed (iid) instances of  $W$ , then  $\mathbb{V}_{\text{rel}}(\bar{W}) = \mathbb{V}_{\text{rel}}(W)/r$ . Let  $\tilde{\epsilon} = (1 + \epsilon)^{1/2} - 1$ . For  $\lceil 2e\sqrt{10}\tilde{\epsilon}^{-2} \rceil$  iid draws of  $W$ ,  $\mathbb{V}_{\text{rel}}(\bar{W}) \leq \tilde{\epsilon}^{-2}/10$ .

Chebyshev's Inequality says that for a random variable  $X$  with finite relative variance:  $\mathbb{P}((1 - \epsilon)\mathbb{E}[X] \leq X \leq (1 + \epsilon)\mathbb{E}[X]) \geq 1 - \mathbb{V}_{\text{rel}}(X)\epsilon^2$ . Hence

$$\mathbb{P}((1 + \tilde{\epsilon})^{-1}\mathbb{E}[W] \leq \bar{W} \leq (1 + \tilde{\epsilon})\mathbb{E}[W]) \geq 1 - 1/10.$$

Similarly,  $\mathbb{P}((1 + \tilde{\epsilon})^{-1}\mathbb{E}[V] \leq \bar{V} \leq (1 + \tilde{\epsilon})\mathbb{E}[V]) \geq 1 - 1/10$ .

Therefore, the chance that step 1 successfully gives a basic estimate of  $\ln(Z(\beta)/Z(0))$ , step 2 creates a well balanced schedule, and step 3 gives  $\bar{W}$  and  $\bar{V}$  both within a factor of  $(1 + \tilde{\epsilon})$  of their respective means is at least  $1 - 1/100 - 4/100 - 1/10 - 1/10 = 75\%$  by the union bound.

If both  $\bar{W}$  and  $\bar{V}$  are within  $1 + \tilde{\epsilon}$  of their means, then  $\bar{W}/\bar{V}$  is within  $(1 + \tilde{\epsilon})^2 = 1 + \epsilon$  of  $\mathbb{E}[\bar{W}]/\mathbb{E}[\bar{V}] = Z(\beta)/Z(0)$ , completing the proof.  $\square$

3.4. *The running time of the basic algorithm.* How many samples does Algorithm 2.2 take on average?

THEOREM 3.2. *When  $n \geq 4$ , and  $\epsilon \leq 1/10$ , Algorithm 2.2 takes on average at most*

$$(q + 1)[5 + (2 + \ln(2n))(14.9 \ln(100(2 + \ln(2n)))(q + 1)) + 48.2\epsilon^{-2}].$$

*samples. For fixed  $\epsilon$  the number of samples is  $O(q[\ln(n)(\ln(q) + \ln(\ln(n)))])$ .*

PROOF. A run of TPA uses a number of samples that is one plus a Poisson random variable with mean  $z(\beta) - z(0)$ , so on average  $q + 1$  samples. So step 1 takes  $5q + 5$  samples on average. From the concavity of the  $\ln$  function and Jensen's inequality, the second step takes at most

$$\lceil (2/3)(2 + \ln(2n)) \rceil \lceil 22 \ln(100(2 + \ln(2n)))(q + 1) \rceil q$$

samples on average. This is bounded above by

$$q[14.9(2 + \ln(2n)) \ln(100(2 + \ln(2n))(q + 1))].$$

The resulting schedule has on average at most  $q/(d/k) + 1 = (2/3)[2 + \ln(2n)]q + 1$  intervals in it, and so the third step of the algorithm generates a number of sample that (on average) is at most

$$(2e\sqrt{10})(2/3)(2 + \ln(2n))(q + 1)((1 + \epsilon)^{1/2} - 1)^{-2}.$$

When  $\epsilon \leq 1/10$ ,  $(1 + \epsilon)^{1/2} - 1 \geq \epsilon/2.05$ , so the number of samples in this section can be bounded by

$$48.2(2 + \ln(2n))(q + 1)\epsilon^{-2}.$$

□

**4. Variations and extensions.** The previous two sections gave an algorithm for when  $-H(x) \in \{0, 1, 2, \dots, n\}$ . In this section algorithms are given for when  $H(x) \geq 0$ , or for when  $H(x)$  changes sign. The section ends with a discussion of the use of approximate rather than exact samples.

*4.1. Dealing with a nonnegative Hamiltonian function.* When  $H(x) \geq 0$ , then  $Z(\beta)$  is a decreasing function of  $\beta$ . The analysis of the relative variance of  $W$  and  $V$  remains identical, the only thing that changes is the TPA algorithm for obtaining the initial estimate of  $q$  and the well balanced schedule.

**ALGORITHM 4.1.** TPA for Gibbs distributions with  $H(x) \geq 0$  takes as input a value  $\beta > 0$  together with an oracle for generating random samples from  $\pi_b$  for  $b \in [0, \beta]$ , and returns a set of values  $0 < b_1 < b_2 < \dots < b_\ell < \beta$  such that  $\{z(b_1), \dots, z(b_\ell)\}$  forms a Poisson point process of rate 1 on the interval  $[z(0), z(\beta)]$ . It operates as follows

1. Start with  $b$  equal to 0, and  $B$  equal to the empty set.
2. Draw a random sample  $X$  from  $\pi_b$ , and draw  $U$  uniformly from  $[0, 1]$ .
3. Let  $b = b - \ln(U)/H(X)$ , unless  $H(X) = 0$ , in which case set  $b = \infty$ .
4. If  $b < \beta$ , then add  $b$  to the set  $B$ , and go back to step 2.

Everything else about the algorithm remains identical.

4.2. *Dealing with a Hamiltonian that is both positive and negative..* Suppose that  $H(x) \in \{-n, -n+1, \dots, -1, 0, 1, \dots, n\}$ . Then neither of the previous sections apply. However, it is possible to shift the  $H(x)$  function without changing the distributions  $\pi_\beta$ . Consider  $H_{\text{shift}}(x) = H(x) + c$ . Then for  $X \sim \pi_\beta$  with  $H_{\text{shift}}(x)$ :

$$\mathbb{P}(X = x) = \frac{\exp(-\beta H_{\text{shift}}(x))}{Z_{\text{shift}}(\beta)} = \frac{\exp(-\beta c) \exp(-\beta H(x))}{Z_{\text{shift}}(\beta)},$$

so  $Z_{\text{shift}}(\beta) = Z(\beta) \exp(-\beta c)$ . This means that by making  $c = -n$ , it is possible to insure that  $H(x) \leq 0$ . But even better, if  $c = -2n$ , it is possible to insure that  $H(x) \in \{-2n, -2n+1, \dots, -n\}$ .

This means that  $z'(\beta)/z'(0) \leq 2$ , so the extraneous  $\ln(n)$  factor from Theorem 3.2 is removed.

LEMMA 4.1. *For  $H(x) \in \{-2n, -2n+1, \dots, -n\}$ , and  $\eta_i \leq \eta$ ,*

$$\mathbb{V}_{\text{rel}}(W) = \mathbb{V}_{\text{rel}}(V) \leq 2^{\eta/2}.$$

PROOF. Recall  $\mathbb{V}_{\text{rel}}(W) = \mathbb{V}_{\text{rel}}(V) = \prod_i \exp(2\delta_i/n)$ , and  $z'(\beta_{i+1})/z'(\beta_i) \geq \exp(4\delta_i/\eta_i)$ . So

$$\frac{z'(\beta)}{z'(0)} = \frac{z'(\beta_1)}{z'(\beta_0)} \dots \frac{z'(\beta_\ell)}{z'(\beta_{\ell-1})} \geq \prod_i \exp(4\delta_i/\eta_i).$$

Since  $H(x) \in \{-2n, -2n+1, \dots, -n\}$ , so is  $\mathbb{E}[-H(X)]$  for  $X \sim \pi_0$  or  $X \sim \pi_\beta$ , so  $z'(\beta)/z'(0) \leq 2$ . Raising to the  $\eta/2$  power finishes the proof.  $\square$

There is a cost to the shift:  $Z_{\text{shift}}(\beta)/Z_{\text{shift}}(0) = Z(\beta) \exp(-\beta(-2n))/Z(0)$ , so  $q_{\text{shift}} = q + 2n\beta$ . Using  $\eta = 2/\ln(2)$  then bounds the relative variance by  $e$  as before.

THEOREM 4.1. *In Algorithm 2.2 applied to the distribution with the shifted Hamiltonian, let  $\hat{q}_1$  be the size of the Poisson point process created with 5 runs of TPA in step 1. Let  $d = \lceil 22 \ln(200(\ln(2))^{-1}(\hat{q}_1 + 2n\beta + 1)) \rceil$ , and  $k = (2/3) \ln(2)d$ . Then the algorithm output is within  $1+\epsilon$  of  $Z(\beta)/Z(0)$  with probability at least  $3/4$ .*

PROOF. Let  $\eta = 2/\ln(2)$  so that  $\mathbb{V}_{\text{rel}}(W) = \mathbb{V}_{\text{rel}}(V)$  is bounded above by  $e$ , then the rest of the proof is identical to that of Theorem 3.1.  $\square$

THEOREM 4.2. *The average number of Gibbs distribution draws used by Algorithm 2.2 run with the parameters in Theorem 4.1 is bounded above by*

$$(q + 2n\beta + 1)[5 + 10.7 \ln(69.4(q + 2n\beta + 1)) + 16.7\epsilon^{-2}].$$

PROOF. With  $\eta = 2/\ln(2)$ , Lemma 4.1 yields  $\mathbb{V}_{\text{rel}}(V) = \mathbb{V}_{\text{rel}}(W) = e$ . Then the first step of the algorithm uses on average  $5(q + 2n\beta + 1)$  samples, the second step uses on average at most

$$(q_{\text{shift}} + 1) \lceil (4/3)(\ln(2)/2) \lceil 22 \ln(200(\ln(2)/2)(q_{\text{shift}} + 1)) \rceil \rceil$$

steps. The final step then takes on average

$$q_{\text{shift}} \lceil e\sqrt{10}((1 + \epsilon)^{1/2} - 1)^{-2} \rceil (2/3) \ln(2) \leq 16.7 q_{\text{shift}} \epsilon^{-2}$$

draws from the Gibbs distribution. Adding these together then gives the result.  $\square$

4.3. *Dealing with approximate samples.* Until now, the algorithms considered have been analyzed under the assumptions that samples drawn exactly from  $\pi_\beta$  are available. While this is true in many instances (see for example [4, 6, 10]) is not true for all problems. For instance, the only algorithm for generating samples from the Ising model that provably runs in polynomial time is a Monte Carlo Markov chain (MCMC) approach of Jerrum and Sinclair [8] where the samples obtained are arbitrarily close to the  $\pi_\beta$  distribution.

This problem of using approximate instead of exact samples can be dealt with in a simple way using a technique from [13]. Consider the total variation distance between distribution  $\pi$  and  $\tau$ :  $\text{TV}(\pi, \tau) = (1/2) \sum |\pi(\{x\}) - \tau(\{x\})|$ . It is always possible to create a *coupling* [3], a bivariate random variable  $(X, Y)$  such that  $X \sim \pi$ ,  $Y \sim \tau$  and  $\mathbb{P}(X \neq Y) = \text{TV}(\pi, \tau)$ .

By making the value of  $S$  large enough that the probability that the algorithm requires more than  $S$  samples is small, then each approximate sample can be coupled with a perfect sample. The probability of failure can then be bounded (with the union bound) by the probability that the perfect samples fail to approximate the target plus the probability that any of the perfect samples are different from the approximate samples.

## References.

- [1] BEZÁKOVÁ, I., STEFANKOVIC, D., VAZIRANI, V. V. and VIGODA, E. (2008). Accelerating simulated annealing for the permanent and combinatorial counting problems. *SIAM J. Comput.* **37** 1429–1454.
- [2] BROOKS, S., GELMAN, A., JONES, G. and MENG, X., eds. (2011). *Handbook of MCMC*. Chapman & Hall/CRC Press.
- [3] DOEBLIN, W. (1933). Exposé de la théorie des chaînes simples constantes de Markov à un nombre fini d'états. *Rev. Math. de l'Union Interbalkanique* **2** 77–105.
- [4] FILL, J. A. and HUBER, M. L. (2010). Perfect simulation of Vervaat perpetuities. *Electron. J. Probab.* **15** 96–109.

- [5] FISHMAN, G. S. (1994). Choosing sample path length and number of sample paths when starting in the steady state. *Oper. Res. Letters* **16** 209–219.
- [6] HUBER, M. (2004). Perfect Sampling Using Bounding Chains. *Annals of Applied Probability* **14** 734–753.
- [7] HUBER, M. L. and SCHOTT, S. (2010). Using TPA for Bayesian inference. *Bayesian Statistics 9* 257–282.
- [8] JERRUM, M. and SINCLAIR, A. (1993). Polynomial-time approximation algorithms for the Ising model. *SIAM J. Comput.* **22** 1087–1116.
- [9] METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equation of state calculation by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.
- [10] PROPP, J. G. and WILSON, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms* **9** 223–252.
- [11] RESNICK, S. (1992). *Adventures in Stochastic Processes*. Birkhäuser.
- [12] SWENDSEN, R. and WANG, J.-S. (1986). Replica Monte Carlo simulation of spin glasses. *Phys. Rev. Lett.* **57** 2607–2609.
- [13] ŠTEFANKOVIČ, D., VEMPALA, S. and VIGODA, E. (2009). Adaptive Simulated Annealing: A Near-Optimal Connection between Sampling and Counting. *J. of the ACM* **56** 1–36.
- [14] VALLEAU, J. P. and CARD, D. N. (1972). Monte Carlo Estimation of the Free Energy by Multistage Sampling. *J. Chem. Phys.* **57** 5457–5462.

DEPARTMENT OF MATHEMATICAL SCIENCES  
CLAREMONT MCKENNA COLLEGE  
850 COLUMBIA AVENUE  
CLAREMONT, CA 91711  
E-MAIL: [mhuber@cmc.edu](mailto:mhuber@cmc.edu)  
URL: <http://www.cmc.edu/pages/faculty/MHuber/>